International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 12, Number 10 (2019), pp. 1813-1817

Development of a Konkani Language Dataset for Automatic Text Summarization and its Challenges

Jovi D'Silva[1] and Dr. Uzzal Sharma[2]

[1]*Research Scholar, Assam Don Bosco University, Guwahati, Assam, India*

[2]*Assistant Professor, Assam Don Bosco University, Guwahati, Assam, India*

Abstract

Text summarization has gained tremendous popularity in the research field over the last few years. Automatic Text summarization attempts to automate the summarization task, which would otherwise, be done by humans. Research has progressed a lot in the said domain in languages such as English. However, Indian languages such as Hindi are gathering interests of a lot of researchers. India has a wide cultural diversity with 22 major languages and approximately 13 various scripts and, not much research has been made in these languages. The primary concern would be finding a dataset to proceed with solving the problem at hand, which is difficult and also the script poses a major challenge. This paper introduces a Konkani language dataset in the domain of Konkani literature which is written in Devanagari script and also gives an account of the challenges faced during the creation of the dataset.

Keywords: Automatic text summarization, Konkani dataset, Rouge.

I. INTRODUCTION

The volume of the content being added to the internet is growing at an exponential pace. New content is uploaded to the Internet every minute and there is a wide lingual diversity observed in the uploaded content. The large volume of data makes a reader's task more tedious as one has to read through every article to interpret the contents.

Automatic text summarization can aid readers tremendously to sift through the content more efficiently by presenting a summary of a large document to readers.

As mentioned in [1], Automatic Text Summarization is a process by which short and concise computer generated summaries are produced of a given text document that is provided as an input document. Summaries could be generated from a single input document or from multiple input documents. There are numerous text summarizers in English, but when it comes to Indian languages the scope is very limited. However, there has been research progressing in languages like Hindi, Bengali, Tamil, and Kannada as stated in [2].

In this paper, an attempt is made to facilitate research in the field of Automatic-text Summarization, pertaining to Konkani language, which is one of the many languages spoken in India, by compiling a dataset for the task of automatic text summarization. Konkani language has been a relatively unexplored territory in the domain of text summarization, and, considering the rapid growth of content on the internet in the language, there is a need for a tool that can assist in summarization of this content.

Text summarization can be stated as the process of creating short and precise summaries of a long document that convey the meaning of the content in a gist. Automatic Text Summarization methods create summaries that preserve the key elements of the contents of the document without altering the actual meaning [3].

Text summaries are mainly classified under the following two types, as stated in [2], [4]:

- Extractive summary
- Abstractive summary

In extractive summarization, significant phrases and sentences are determined and these key aspects of the text make up the summary [2], [4], [5]. Abstractive summary involves comprehending the contents of the text prior and then constructing a summary that may contain new phrases and sentence formations as compared to the original text [2], [4], [5]. The latter method is a more exigent and popular way in which humans work while preparing summaries.

Various approaches to proceed with the summarization are rule based and machine learning [4]. The machine learning approach further bifurcates into supervised, unsupervised content extraction methods with reference to [4]. In supervised learning, the machine is trained using a dataset consisting of a set of human-annotated articles, whereas, unsupervised learning does not require any data for training.

A lot of work related to datasets in English pertains to news articles, blogs and research abstracts. However, no dataset is available in Konkani that is specifically developed for automatic text summarization. In an attempt to curb this primary hurdle, a Konkani literature dataset has been created and presented in this paper. It gives a detailed account of the dataset generation process along with the challenges faced while achieving this task. The novelty of the dataset is that it intends to capture the domain of Konkani literature by using actual works of literature, such as short stories rather than using news articles, blogs and research abstracts which has already been heavily worked upon.

II. RELATED WORK

Most of the datasets available at present are in English, as it is one of the most widely spoken languages in the world and also most extensively used language on the internet. Authors of [6] have presented a gold corpus of 200 English short stories. Summaries are generated by picking short text segments spread across an article. The summaries generated are similar to telegrams, and hence, termed as Telegraphic Summarization by the authors. The authors of [7] present "The Opinosis" dataset which consists of 51 articles. The articles describe a product's features and have a compilation of the customer reviews that purchased a particular product. Every article in the dataset has 5 "gold" summaries written manually.

The National Institute of Standards and Technology (NIST) has initiated evaluation in the sphere of text summarization called The Document Understanding Conference (DUC) for further progress in the summarization expanse, and, enable researchers to participate and contribute in large-scale experiments in the related domain [8]. NIST formed 60 reference sets, every set comprised of documents, single-document abstracts and multi-document abstracts/extracts.DUC has several datasets, namely, DUC 2001 to DUC 2007. However, DUC 2002 is mainly composed of news articles compiled from various newspapers.

"The Gigaword English Corpus" [9] is another dataset in the English language that is made of comprehensive chronicles of newswire text data that has been gathered over a couple of years by the Linguistic Data Consortium (LDC). The authors of [10] propose an effective method of automatically collecting huge volumes of news-related multi-document summaries in correspondence to the reactions of the social media. A linked document cluster is further synthesized to construct a reference summary which can encompass maximum key points from articles. Using the assembled data as a training resource, the feat of the summarizer showed improvement on all test sets.

"SciSumm" corpus contains a release of the corpus of scientific document summarization and annotations from the WING NUS group [11]. It consists of manually annotated training set of 40 articles and citing papers. Authors of [12] present a way of collecting corpora for automatic live blog summarization.

"TutorialBank" is an openly accessible dataset which was produced with the intent to promote Natural Language Processing (NLP) learning and exploration [13]. There are over 6,300 manually gathered and classified resources on NLP as well as resources from relevant domains of computer science. This dataset is the greatest manually selected resource corpus intended for further learning in NLP. The authors of [14] present a dataset that comprises of Australian legal cases brought by the Federal Court of Australia (FCA). It is textual corpus of 4000 legal cases generated with the intent to facilitate research in automatic summarization and citation analysis.

"TIPSTER" Text Summarization Evaluation Conference (SUMMAC) released a compilation of 183 documents from the Computation and Language collection (cmp-lg) that has been made available to the information retrieval, extraction and summarization researchers as a general resource [15]. It is a collection of scientific paper documents that emerged in sponsored conferences of the Association for Computational Linguistics (ACL).

News Summary is a dataset compiled from summarized news articles [16]. It consists of 4515 examples and also consists of Author name, Headlines, URL of Article, Short text, Complete Article. The BBC News Summary dataset was built for extractive text summarization and consists of 417 political news related articles of BBC from the year 2004 to the year 2005 [17].

"Sentence-compression" [18] is a large collection of uncompressed and compressed news articles' sentences. "The Columbia Summarization Corpus" (CSC)[19] has a sum total of 166,435 summaries containing 2.5 million sentences and covering 2,129 days in the time period from 2003 to 2011. These were collected from an online news summarization system called "Newsblaster" that seeks various news articles over the web.

"WikiHow" [20] is another Large Scale Text Summarization Dataset. Each article in this dataset consists of multiple paragraphs and every paragraph begins with a sentence summarizing it. It is constructed with the help of content from "WikiHow" website.

Other languages too are taking over the world of web, and, datasets have been created specifically for automatic text summarization in languages that are quite divergent syntactically and structurally in comparison to English. Once such language is Arabic, [21] presents one such dataset with the challenges and various metrics that the authors used to evaluate the work. A Chinese dataset for the purpose of text summarization is presented in [22], which is the largest Chinese dataset and is created by compiling data from a Chinese micro-blogging site.

"The PUCES corpus of Automatic Text Summarization" [23] is a French language dataset created for automatic text summarization. It consists of a set of 221 summaries constructed by 17 systems, human annotators and a set of 132 human generated abstracts. The source documents have 30 sentences referring to 2 topics: "A new electronic microchip for computers" and "Fleas invasion in a military company".
Indian languages like Bengali, Gujarati, Tamil, Hindi and Malayalam have also gained interests of several researchers in the area of text summarization [2], [24]. These languages are drastically different from English in terms of language structure and also script. Similar work with respect to automatic text summarization in Hindi is presented in [25]. It uses supervised machine learning approach using Hindi news articles' dataset.

Most of the previously built datasets in various languages are based on newspaper articles, research abstracts and blogs. There is no dataset available in Konkani language, let alone one that covers the domain of Konkani literature. This dataset consists of folk tales written in the Konkani language by prominent Konkani authors. Folk tales have strong historical and cultural ties and are usually passed down generation to generation mainly through storytelling. Folk tales also include legends, myths and fairytales.

III. CONSTRUCTING KONKANI DATASET

The source text for generating summaries is extracted from Konkani books written by various authors from Goa. These books contain many folk tales from Goa, Russia and other parts of India written in Devanagari script. Thus, each document in the dataset is a summary of a folk tale from any of the above books. So, the dataset is composed of a total of 71 folk tales from 5 books.

The Konkani folktale books are extremely rare and difficult to obtain. The books were carefully chosen based on the content and availability.

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 12, Number 10 (2019), pp. 1813-1817

Two human summarizers, proficient in Konkani, read each folk tale and generated an abstractive summary which was not more than 300 words in length. Each human summarizer worked independently of the other to ensure that one's work did not influence the other.

Fig. 1 gives a visual understanding of the dataset generation process. Two human summarizers generate two sets of summaries obtained from the same set of source documents. The main reason behind having two human summarizers is to measure the accuracy of the automatic text summarizer against two sets of human-generated summaries.

In addition, each of the folk tales was annotated. This is annotation was done by a human so as to indicate which sentences are good candidates for inclusion in a summary and which are not. This annotation is done so as to facilitate the training of a supervised learning algorithm. Unsupervised algorithms do not need any such training data.

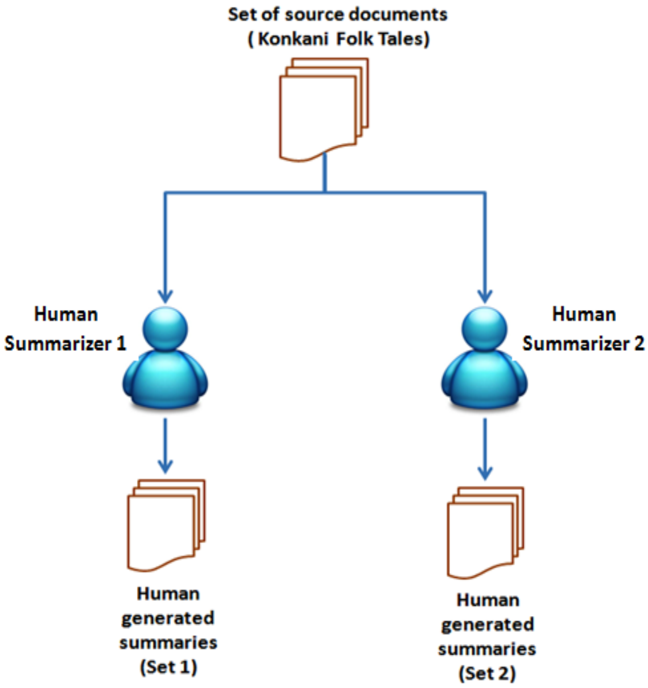

Fig. 1. Dataset Preparation Process

IV. EVALUATION

The Primary purpose of creating a Konkani literature dataset was for detailed research on Automatic Text Summarization. The primary idea of the author is constructing an automated tool for producing text summaries of Konkani documents.

The two human generated summaries that are different from each other serve as reference summary. Thereafter, the quality of the auto-generated summary can be measured against the human generated summaries to check for coherence and its ability to convey the important points of the input document in a clear and concise manner. An effective tool that can aid in achieving this purpose is the ROUGE package which is a short form of "Recall-Oriented Understudy for Gisting Evaluation". ROUGE is one of the ways to compute the effectiveness of auto generated summaries.

The author of [26] has presented the ROUGE package and demonstrated the different measures used to test the effectiveness and the quality of a summary. The different measures used in the package determine the summary quality by comparison of the machine generated summary with the "ideal" human generated summaries.

The four measures used in the package viz., ROUGE-N, ROUGE-L, ROUGE-W and ROUGE-S along with their evaluation procedures have been presented by the author of [26]. The evaluation of the scores obtained after using the package gives a clear insight on the quality of the system generated summaries. We come across the definitions of ROUGE-N, ROUGE-L and ROUGE-S in [27], [28].

ROUGE 2.0 is a Java toolkit that can be used for automatic summary evaluation tasks [27]. It utilizes the ROUGE metrics that work by correlating an automatically composed summary against a set of reference summaries which are generally human-generated. With reference to ROUGE, Precision and Recall are two very significant measures that quantitatively indicate a good summary using overlap [27], [28]. Recall indicates how much of the human-generated summary is actually being captured by the automatically-generated summary. Recall can be computed as denoted by (1),

$$Recall = \frac{number\_of\_overlapping\_words}{total\_words\_in\_humangenerated\_summary} \qquad (1)$$

Precision is defined as how much of the automatically-generated summary is actually relevant or needed, as shown in (2).

$$Precision = \frac{number\_of\_overlapping\_words}{total\_words\_in\_autogenerated\_summary} \qquad (2)$$

V. CHALLENGES

The primary challenge while constructing this dataset was the lack of availability of books on Konkani literature since the language is not very popular amongst writers in comparison with languages like English. There are few rare books that were written a long time ago, but in modern times there are very few writers choosing to write in Konkani and so not many resources are available.

The number of writers in Konkani are also very few in number than that of languages like Hindi, Marathi and the like. The state of Goa where the language is primarily spoken is very small in terms of the area and has limited population that speaks Konkani in comparison to the population speaking the other popular languages like Hindi. There are some places adjoining Goa where people speaking various dialects of Konkani are encountered. These dialects often have a mixture of other languages in them, and, for the sake of adhering to one unique dialect, the authors have attempted to procure Konkani literature that was primarily written by authors speaking the dialect of the language unique to the state of Goa.

Another major challenge encountered was while seeking summarizers for human summary generation. The problems faced were the similar to that faced when finding Konkani literature as mentioned above. Out of the limited Konkani Speaking population (speaking the required dialect), there are very few people who have chosen to major in this language and have grammatical knowledge in the subject. It was indeed a cumbersome task to find qualified people who were willing to contribute towards this work.

Another major challenge is the lack of available text processing tools such as tokenizers, stemmers, taggers and parsers for the Konkani language. Due to this most of the sentence segmentation had to be done manually as existing tools were not available or did not work as desired.

VI. CONCLUSION

We have constructed a new dataset for the purpose of automatic text summarization in Konkani language. This is a dataset from the domain of literature consisting of folktales written in Konkani language.

Text summarization is mostly limited to some well-researched languages and there is a need for this research to be extended to other languages as well. In addition, research should also be considered in domains such as folk tales, short stories rather than just news articles, blogs and research abstracts.

ACKNOWLEDGEMENTS

We would like to express heartfelt appreciation to the authors of the Konkani books who have contributed towards building the Konkani dataset presented in this paper. We wish to thank Dr. Jayanti Naik, Ms. Kamaladevi Rao Deshpande and family, Ms. Maya Kharangate for providing us with their literary works. We also like to thank the publishers Yugved Prakashan, Rajaee Prakashan and Asmitai Pratishthan.

We would like to thank the human summarizers, Ms. Shubha Barad, Ms. Teffany Gama and Ms. Disha Mashelkar, and lastly, Mr. Rohan Kerkar for his valuable time and expertise in formatting the dataset.

REFERENCES

[1] Patil A, Dalmia S et al. Automatic Text Summarizer. 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI). New Delhi, India. Sept. 2014.

[2] Gupta V. A Survey of Text Summarizers for Indian Languages and Comparison of their Performance. Journal of Emerging Technologies in Web Intelligence. November 2013. Vol. 5. No. 4. pp. 361-366.

[3] Allahyari M, Pouriyeh S, Assef M, Safaei S, Trippe ED, Gutierrez JB, Kochut K. Text Summarization Techniques: A Brief Survey. arXiv:1707.02268. USA. July 2017.

[4] Nimavat K, Joshiara HA. Query-Based Summarization Methods for Conversational Agents: An Overview. International Journal of Advanced Research in Computer Science. Volume 8. No. 8. ISSN No. 0976-5697. September-October 2017.

[5] Verberne S, Krahmer E, Hendrickx I et al. Creating a Reference Data Set for the Summarization of Discussion Forum Threads. Lang Resources & Evaluation (2018). Springer Netherlands. 52: 461. 2018. Available: https://doi.org/10.1007/s10579-017-9389-4.

[6] Malireddy C, Somisetty SNM, Shrivastava M. Gold Corpus for Telegraphic Summarization. Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing. Santa Fe, New Mexico. August 2018. pp. 71–77.

[7] Ganesan K, Zhai CX, Han J. Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions. Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). Coling 2010 Organizing Committee. Beijing, China. August 2010. pp. 340–348.

[8] Document Understanding Conferences (DUC). National Institute of Standards and Technology (NIST). Available: https://www-nlpir.nist.gov/projects/duc/guidelines/2002.html.

[9] Graff D, Cieri C. English Gigaword LDC2003T05. Web Download. Philadelphia: Linguistic Data Consortium, 2003. ISBN 1-58563-260-0.

[10] Cao Z, Chen C, Li W, Li S, Wei F, Zhou M. TGSum: Build Tweet Guided Multi-Document Summarization Dataset. arXiv:1511.08417, 2015.

[11] Jaidka K, Chandrasekaran MK, Rustagi S, Kan M. Overview of the CL-SciSumm 2016 Shared Task. In Proceedings of Joint Workshop on Bibliometric-enhanced Information Retrieval and NLP for Digital Libraries (BIRNDL 2016). Newark, NJ, USA. June 23, 2016, pp. 93-102.

[12] Avinesh PVS, Peyrard M, Meyer CM. Live Blog Corpus for Summarization. Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC). Miyazaki, Japan. May 2018. pp. 3197-3203. Available: http://www.lrec-conf.org/proceedings/lrec2018/pdf/317.pdf

[13] Fabbri AR, Li I, Trairatvorakul P, He Y, Ting WT, Tung R, Westerfield C, Radev DR. TutorialBank: A Manually-Collected Corpus for Prerequisite Chains, Survey Extraction and Resource Recommendation. 2018, 10.18653/v1/P18-1057, pp. 611-620.

[14] Galgani F, Compton P, Hoffmann A. Towards Automatic Generation of Catchphrases for Legal Case Reports. 13th International Conference on Intelligent Text Processing and Computational Linguistics. Lecture Notes in Computer Science. Springer Berlin Heidelberg. 2012. Vol. 7182. pp. 414-425.

[15] Mani I, House D, Klein G et al. TIPSTER Text Summarization Evaluation Conference (SUMMAC). May 1998. Available: https://www.nlpir.nist.gov/related_projects/tipster_summac/index.html

[16] News Summary. Available: https://www.kaggle.com/sunnysai12345/news-summary

[17] BBC News Summary. Available:https://www.kaggle.com/pariza/bbc-news-summary/data

[18] Filippova K, Altun Y. Overcoming the Lack of Parallel Data in Sentence Compression. Proceedings of the

2013 Conference on Empirical Methods in Natural Language Processing (EMNLP '13). Seattle, Washington, USA. 18-21 October 2013. pp. 1481-1491.

[19] Wang WY, Thadani K, McKeown KR. Identifying Event Descriptions using Co-training with Online News Summaries. Proceedings of 5th International Joint Conference on Natural Language Processing. Chiang Mai, Thailand. November 2011. pp. 282–291.

[20] Koupaee M, Wang WY. WikiHow: A Large Scale Text Summarization Dataset. Oct 2018.

[21] Al Qassem LM, Wang D, Al Mahmoud Z, Barada H, Al-Rubaia A, Almoosa NI. Automatic Arabic Summarization: A survey of methodologies and systems. 3rd International Conference on Arabic Computational Linguistics, ACLing 2017. Procedia Computer Science. Dubai, UAE. 5-6 November 2017. Vol. 117. pp. 10-18.

[22] Hu B, Chen Q, Zhu F. LCSTS: A Large Scale Chinese Short Text Summarization Dataset. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal. 17-21 September 2015. pp. 1967–1972.

[23] Cabrera-Diego LA, Torres-Moreno JM, Durette B. Evaluating Multiple Summaries Without Human Models: A First Experiment with a Trivergent Model. International Conference on Applications of Natural Language to Information Systems Natural Language Processing and Information Systems. NLDB 2016. Springer Cham. Vol. 9612. pp. 91-101.

[24] Dhanya PM, Jathavedan M. Comparative Study of Text Summarization in Indian Languages. International Journal of Computer Applications. August 2013. Vol. 75– No.6. pp. 17-21.

[25] Desai N, Shah P. Automatic Text Summarization Using Supervised Machine Learning Technique For Hindi Language. IJRET: International Journal of Research in Engineering and Technology. Jun-2016. eISSN: 2319-1163. pISSN: 2321-7308. Volume: 05 Issue: 06.

[26] Lin CY. ROUGE: A Package for Automatic Evaluation of summaries. Publisher: Association for Computational Linguistics. Barcelona, Spain. 2004. pp. 74-81.

[27] Rxnlp. ROUGE. Available: http://rxnlp.com/how-rouge-works-for-evaluation-of-summarization-tasks/#.XPyyfxYzbIU

[28] Ganesan K. What is ROUGE and how it works for evaluation of summaries?. Available: http://kavita-ganesan.com/what-is-rouge-and-how-it-works-for-evaluation-of-summaries/#.W5LhLJNKidt/